LUMS

A Not-for-Profit University

Department of Electrical Engineering
School of Science and Engineering

EE514/CS535 Machine Learning

HOMEWORK 2 – SOLUTIONS

Due Date: 23:55, Friday, April 15, 2022 (Submit online on LMS)
Format: ?? problems, for a total of 100 marks
Contribution to Final Percentage: 2.5%
Instructions:

- Each student must submit his/her own hand-written assignment, scanned in a single PDF document.
- You are allowed to collaborate with your peers but copying your colleague's solution is strictly prohibited. Anybody found guilty would be subjected to disciplinary action in accordance with the university rules and regulations.
- Note: Vectors are written in lowercase and bold in the homework. For your written submission, you may use an underline instead. In addition, use capital letters for matrices and lowercase for scalars.

Problem 1 (10 marks)

Bayes Theorem - Bayesian machine learning is a paradigm for constructing statistical models based on Bayes' Theorem. To get more insight into the significance of bayes theorem, try to solve this problem.
Suppose you have symptoms for COVID-19, and you decide to take a diagnostic test to determine if you are infected or not. You being a skeptical individual question the reliability of the outcome of the test and find out that the probability of a positive test given that a person is actually infected (**sensitivity**) is 87.5% and the probability of a negative test given that a person is not infected (**specificity**) is 97.5%. You also know that the prevalence of the infection in your vicinity is 10%. Now suppose that the diagnostic test turns out positive. In this case, what is the probability that you are actually infected with the virus?

Solution: Let $H = 1$ be the event that you are infected, and $H = 0$ be the event you are not infected. Let $Y = 1$ if the test is positive, and Y = 0 if the test is negative. We want to compute $p(H = h|Y = y)$, for $h = 1$, where $y = 1$ is the observed test outcome.
We can think of this as a form of binary classification, where H is the unknown class label, and y is the feature vector.

$$\begin{aligned}
p(H=1|Y=1) &= \frac{p(Y=1|H=1)p(H=1)}{p(Y=1|H=1)p(H=1)+p(Y=1|H=0)p(H=0)} \qquad (1)\\
&= \frac{TPR \times prior}{TPR \times prior + FPR \times (1-prior)} \qquad (2)\\
&= \frac{0.875 \times 0.1}{0.875 \times 0.1 + 0.025 \times 0.9} \qquad (3)\\
&= 0.795 \qquad (4)
\end{aligned}$$

So there is a 79.5% chance you are infected.

Note: TPR and FPR are the sensitivity and specificity respectively.

Problem 2 (20 marks)

MAP Estimation - Alice tries to send Bob message M over a communication channel which adds a gaussian noise to the message. The message M is either a 1 or 0 with probability 0.3 and 0.7 respectively. The Gaussian noise is found to be zero mean and unit variance ($N \sim \mathcal{N}(0,1)$). At the receiver, Bob receives a message the message Y.

$$Y = M + N$$

Given that Bob receives a message Y = 0.6, he wants to know whether a 0 or 1 was sent. Using MAP estimate, find whether a 1 or 0 was transmitted.
Note: Gaussian distribution is given by the equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Hint: Try to find the distribution of Y using M and N. Moreover, M is only a constant bias of either 0 or 1. Whenever a constant is added to a Gaussian random variable only affects the mean of the resulting random variable.

Solution: Use MAP estimation to maxamize the posterior probability of p(M = m|Y= 0.7) where m is either 0 or 1.

$$\begin{aligned} \hat{m} &= \underset{m}{\text{argmax}}\ P(M = m|Y = y) && (5)\\ &= \underset{m}{\text{argmax}}\ \frac{f_{Y|M}(y|m)\, P(M = m)}{f_Y(y)} && (6)\\ &= \underset{m}{\text{argmax}}\ f_{Y|M}(y|m)\, P(M = m) && (7) \end{aligned}$$

Given that M=0, Y becomes equal to the noise M ($Y \sim \mathcal{N}(0,1)$), and therefore

$$\begin{aligned} f_{Y|M}(y = 0.6|m = 0) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{0.6^2}{2}\right)\\ &= 0.3332 \end{aligned}$$

Given that M=1, Y becomes Y=N+1 , which is just N but "displaced" by 1 unit. In other words, Y is now a Gaussian random variable with the same variance as N but with mean 1, thus

$$\begin{aligned} f_{Y|M}(y = 0.6|m = 1) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-1)^2}{2}\right)\\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(0.6-1)^2}{2}\right)\\ &= 0.3683 \end{aligned}$$

Using (3), probability that Alice sent 0 given Bob received 0.6 is

$$\begin{aligned} P(M = 0|Y = 0.6) &= f_{Y|M}(0.6|0)\, P(M = 0)\\ &= 0.3332 \times 0.7\\ &= 0.2332 \end{aligned}$$

Again using (3), probability that Alice sent 1 given Bob received 0.6 is

$$\begin{aligned} P(M = 1|Y = 0.6) &= f_{Y|M}(0.6|1)\, P(M = 1)\\ &= 0.3683 \times 0.3\\ &= 0.1105 \end{aligned}$$

Therefore, Alice sent a zero ($\hat{m} = 0$) as this maximizes the posterior probability.

Problem 3 (10 marks)

(a) [**5 marks**] For a softmax function, show that

$$\mathrm{softmax}(\mathbf{a}) = \mathrm{softmax}(\mathbf{a} + b)$$

where **a** is an input vector and b is a scalar. Here we are adding the value b to every dimension of **a**.

Solution:

$$\big(\mathrm{softmax}(\mathbf{a} + b)\big)_i = \frac{\exp(a_i + b)}{\sum_j^d \exp(a_j + b)} == \frac{\exp(a_i)}{\sum_j^d \exp(a_j)} = \big(\mathrm{softmax}(\mathbf{a} + b)\big)_i$$

(b) [**5 marks**] For the following matrix, calculate the softmax for each row.

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 2 \\ 0 & 1 & 2 \\ 1 & 5 & 3 \end{pmatrix}.$$

Solution:

$$\mathrm{softmax}(\mathbf{A}) = \begin{pmatrix} 0.6652 & 0.0900 & 0.2447 \\ 0.0900 & 0.2447 & 0.6652 \\ 0.0159 & 0.8668 & 0.1173 \end{pmatrix}.$$

Problem 4 (20 marks)

We often use regularization to reduce overfitting. In the case of ridge regression, we added

$$\lambda||\mathbf{w}||_2^2 \tag{8}$$

as the regularization term in the objective function to be minimized. In this question, we extend this to logistic regression.

(a) **[5 marks]** Formulate a loss function for the logistic regression and add the regularization term in the objective function.

Solution: Without regularization, we have a loss function

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x_i})\right),$$

where we assume that the true labels are $\{-1, +1\}$. The loss function with regularization term is given by Without regularization, we have a loss function

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x_i})\right) \; + \; \lambda||\mathbf{w}||_2^2$$

(b) **[15 marks]** Another way to minimize the objective function with regularization term is to obtain Maximum A Posteriori (MAP) estimate.

$$\mathbf{w}_{(\text{MAP})} = argmax \prod_{i=1}^{n} P(y_i|x_i, \mathbf{w})P(\mathbf{w}) \tag{9}$$

We make the following assumptions:

$P(y_i|\mathbf{x}_i, \mathbf{w}) = 1/(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)$ for all $i \in 1, 2, \ldots\ldots N$, $\mathbf{w}$ (Prior on $\mathbf{w}$) is normally distributed, which has zero mean and its covariance matrix is a multiple of the identity matrix.

$$P(\mathbf{w}) = \prod_{j=1}^{d} 1/(\sqrt{(2\pi\sigma)})exp(-w_j^2/2\sigma^2) \tag{10}$$

We require you to show that for a particular value of $\lambda$and σ, the MAP estimate is the same as the $\mathbf{w}$ obtained by minimizing the objective function formulated for regularized logistic regression.

Solution: We have MAP estimate given by

$$\begin{aligned}
\mathbf{w}_{(\mathrm{MAP})} &= argmax \prod_{i=1}^{n} P(y_i|x_i, \mathbf{w})P(\mathbf{w}) \\
&= argmax \log \prod_{i=1}^{n} P(y_i|x_i, \mathbf{w})P(\mathbf{w}) \\
&= argmax \sum_{i=1}^{n} \log \Big(P(y_i|x_i, \mathbf{w})P(\mathbf{w}) \Big) \\
&= argmax \sum_{i=1}^{n} \log P(y_i|x_i, \mathbf{w}) \ + \ \log P(\mathbf{w}) \\
&= argmax \sum_{i=1}^{n} \log \frac{1}{1+\exp(-y_i \mathbf{w}^T \mathbf{x_i})} \ + \ \log P(\mathbf{w}) \\
&= argmax \sum_{i=1}^{n} - \log \big(1+\exp(-y_i \mathbf{w}^T \mathbf{x_i})\big) \ + \ \log P(\mathbf{w})
\end{aligned}$$

Substituting $P(\mathbf{w})$ yields

$$\mathbf{w}_{(\mathrm{MAP})} = = argmax - \left(\sum_{i=1}^{n} \log \big(1+\exp(-y_i \mathbf{w}^T \mathbf{x_i})\big) + \frac{\|\mathbf{w}\|_2^2}{2\sigma^2} \right),$$

which is equivalent to minimizing the loss function given in part (a) with $\lambda = \frac{1}{2\sigma^2}$.

Problem 5 (25 marks)

You have been given miniature training and test documents from the actual dataset of movie reviews. The documents belong to either positive, negative, or neutral class.

Sentiment	Text	
Training	Positive	great acting by everyone and amazing movie
	Positive	superb plot and cinematography
	Neutral	average acting but storyline is good
	Negative	lacks proper plot
	Negative	the movie is an utter disaster
Test	?	great acting by Leonardo and amazing storyline

Table 1: Review data

You need to develop a multinomial Naïve Bayes' classifier for this problem by following the steps given below.

(a) [**4 marks**] A list of stop words is given to you

Stop words = [as, if, at, by, and, the, an, but, is]

Apply preprocessing to the training data and test data by removing stop words from them and show the documents after preprocessing.

Solution: Training and test documents after removing stop words
Documents
great acting everyone amazing movie
superb plot cinematography
lacks proper plot
movie utter disaster
average acting storyline good
great acting Leonardo amazing plot

(b) [**3 marks**] You need to work with the preprocessed documents from now onwards. Construct vocabulary from the data and tell its size.

Solution: $V = \{$ great, acting, everyone, amazing, movie, superb, plot, cinematography, lacks, proper, utter, disaster, average, storyline, good $\}$ with $|V| = 15$.

(c) [**5 marks**] Compute prior probabilities.

Solution:

P(positive) = 2/5
P(negative) = 2/5
P(neutral) = 1/5

(d) [**8 marks**] Compute likelihoods of all the words in the training data using Laplace add-one smoothing.

Solution:

For positive class:

P(great — positive) = (1 + 1) / (8 + 15)
P(acting — positive) = (1 + 1) / (8 + 15)

P(everyone — positive) = (1 + 1) / (8 + 15)
P(amazing — positive) = (1 + 1) / (8 + 15)
P(movie — positive) = (1 + 1) / (8 + 15)
P(superb — positive) = (1 + 1) / (8 + 15)
P(plot — positive) = (1 + 1) / (8 + 15)
P(cinematography — positive) = (1 + 1) / (8 + 15)
P(lacks — positive) = (0 + 1) / (8 + 15)
P(proper — positive) = (0 + 1) / (8 + 15)
P(utter — positive) = (0 + 1) / (8 + 15)
P(disaster — positive) = (0 + 1) / (8 + 15)
P(average — positive) = (0 + 1) / (8 + 15)
P(storyline — positive) = (0 + 1) / (8 + 15)
P(good — positive) = (0 + 1) / (8 + 15)

For negative class:

P(great — negative) = (0 + 1) / (6 + 15)
P(acting — negative) = (0 + 1) / (6 + 15)
P(everyone— negative) = (0 + 1) / (6 + 15)

P(amazing— negative) = (0 + 1) / (6 + 15)
P(movie — negative) = (1 + 1) / (6 + 15)
P(superb — negative) = (0 + 1) / (6 + 15)
P(plot — negative) = (1 + 1) / (6 + 15)
P(cinematography — negative= (0 + 1) / (6 + 15))
P(lacks — negative) = (1 + 1) / (6 + 15)
P(proper — negative) = (1 + 1) / (6 + 15)
P(utter — negative) = (1 + 1) / (6 + 15)
P(disaster — negative) = (1 + 1) / (6 + 15)
P(average — negative) = (0 + 1) / (6 + 15)
P(storyline — negative) = (0 + 1) / (6 + 15)
P(good — negative) = (0 + 1) / (6 + 15)

For neutral class:

P(great — neutral) = (0 + 1) / (4 + 15)
P(acting — neutral) = (1 + 1) / (4 + 15)
P(everyone — neutral) = (0 + 1) / (4 + 15)
P(amazing — neutral) = (0 + 1) / (4 + 15)
P(movie — neutral) = (0 + 1) / (4 + 15)
P(superb — neutral) = (0 + 1) / (4 + 15)
P(plot — neutral) = (0 + 1) / (4 + 15)
P(cinematography — neutral) = (0 + 1) / (4 + 15)
P(lacks — neutral) = (0 + 1) / (4 + 15)
P(proper — neutral) = (0 + 1) / (4 + 15)
P(utter — neutral) = (0 + 1) / (4 + 15)
P(disaster — neutral) = (0 + 1) / (4 + 15)
P(average — neutral) = (1 + 1) / (4 + 15)
P(storyline — neutral) = (1 + 1) / (4 + 15)
P(good — neutral) = (1 + 1) / (4 + 15)

(e) **[5 marks]** Now, predict the sentiment of the test data and show your working.

Solution:

P(positive—doc) = P(doc—positive) × P(positive)
= P(great—positive)× P(acting—positive)×P(amazing—positive)×P(plot—positive)×P(positive)
$= ((2 \times 2 \times 2 \times 2)/(8+15)^4) \times (2/5) = 2.28 \times 10^{-5}$

P(negative—doc) = P(doc—negative) × P(negative)
= P(great—negative)×P(acting—negative)×P(amazing—negative)×P(plot—negative)×P(negative)
$= ((1 \times 1 \times 1 \times 2)/(6+15)^4) \times (2/5) = 4.11 \times 10^{-6}$

P(neutral—doc) = P(doc—neutral) × P(neutral)
=P(great—neutral)×P(acting—neutral)×P(amazing—neutral)×P(plot—neutral)×P(neutral)
$= ((1 \times 1 \times 2 \times 1)/(4+15)^4) \times (1/5) = 6.13 \times 10^{-64}$

Hence, the test document belongs to positive class.

Problem 6 (15 marks)

(a) **[5 marks]** Draw a Bayesian network representing Naïve Bayes classifier. Assume that the d features are denoted by $x^{(1)}, x^{(2)}, \ldots, x^{(d)}$ and a class label is denoted by y.

(b) **[10 marks]** Consider a following Bayesian network. We assume that variables are boolean. Write the joint probability distribution $P(A, B, C, D, E, F, G, H, I)$ as a product of conditional distributions factored according to the Bayesian network.

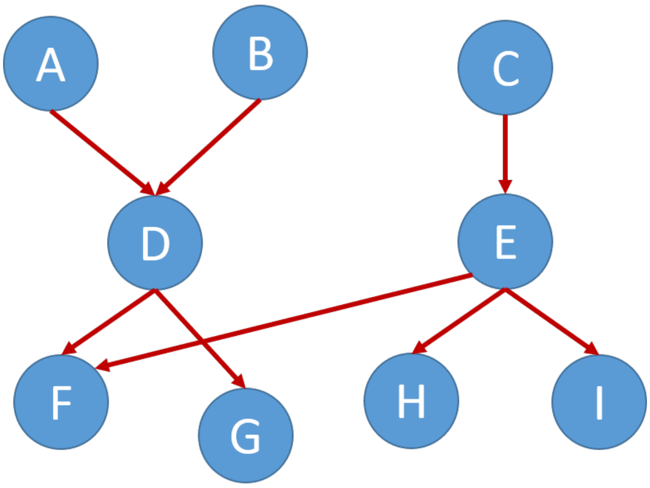

Figure 1: Bayesian Network

Solution:

$$P(A, B, C, D, E, F, G, H, I) = P(A)\,P(B)\,P(C)\,P(D|A, B)\,P(E|C)\,P(F|D, E)\,P(G|D)\,P(H|E)\,P(I|E)$$

— End of Homework —